

A FIXED SIZE STORAGE $O(n^3)$ TIME COMPLEXITY LEARNING ALGORITHM FOR FULLY RECURRENT CONTINUALLY RUNNING NETWORKS

(*Neural Computation*, 4(2):243–248, 1992)

Jürgen Schmidhuber*
Institut für Informatik
Technische Universität München
Arcisstr. 21, 8000 München 2, Germany

Abstract

The RTRL algorithm for fully recurrent continually running networks (Robinson and Fallside, 1987)(Williams and Zipser, 1989) requires $O(n^4)$ computations per time step, where n is the number of non-input units. I describe a method suited for on-line learning which computes exactly the same gradient and requires fixed-size storage of the same order but has an average time complexity¹ per time step of $O(n^3)$.

INTRODUCTION

There are two basic methods for performing steepest descent in fully recurrent networks with n non-input units and $m = O(n)$ input units. ‘Back propagation through time’ (BPTT) (e.g. (Williams and Peng, 1990)) requires potentially unlimited storage in proportion to the length of the longest training sequence but needs only $O(n^2)$ computations per time step. BPTT is the method of choice if training sequences are known to have less than $O(n)$ time steps. For training sequences involving many more time steps than n , for training sequences of unknown length, and for on-line learning in general one would like to have an algorithm with upper bounds for storage and for computations required per

*Current address: Dept. of Computer Science, University of Colorado, Campus Box 430, Boulder, CO 80309, USA, yirgan@cs.colorado.edu

¹Since the acceptance of this paper for publication it has come to my attention that the same algorithm was derived by Ron Williams (Williams, 1989; Williams and Zipser, 1992).

time step. Such an algorithm is the RTRL-algorithm (Robinson and Fallside, 1987)(Williams and Zipser, 1989). It requires only fixed-size storage of the order $O(n^3)$ but is computationally expensive: It requires $O(n^4)$ operations per time step². The algorithm described herein requires $O(n^3)$ storage, too. Every $O(n)$ time steps it requires $O(n^4)$ operations, but on all other time steps it requires only $O(n^2)$ operations. This cuts the average time complexity per time step to $O(n^3)$.

THE ALGORITHM

The notation will be similar to the notation of (Williams and Peng, 1990). U is the set of indices k such that at the discrete time step t the quantity $x_k(t)$ is the output of a non-input unit k in the network. I is the set of indices k such that $x_k(t)$ is an external input for input unit k at time t . $T(t)$ denotes the set of indices $k \in U$ for which there exists a specified target value $d_k(t)$ at time t . Each input unit has a directed connection to each non-input unit. Each non-input unit has a directed connection to each non-input unit. The weight of the connection from unit j to unit i is denoted by w_{ij} . To distinguish between different ‘instances’ of w_{ij} at different times, we let $w_{ij}(t)$ denote a variable for the weight of the connection from unit j to unit i at time t . This is just for notational convenience: $w_{ij}(t) = w_{ij}$ for all t to be considered. One way to visualize the $w_{ij}(t)$ is to consider them as weights of connections to the t -th non-input layer of a feed-forward network constructed by ‘unfolding’ the recurrent network in time (e.g. (Williams and Peng, 1990)). A training sequence with $s + 1$ time steps starts at time step 0 and ends at time step s . The algorithm below is of interest if $s \gg n$ (otherwise it is preferable to use BPTT).

For $k \in U$ we define

$$net_k(0) = 0, \quad \forall t \geq 0 : x_k(t) = f_k(net_k(t)), \quad \forall t > 0 : net_k(t+1) = \sum_{l \in U \cup I} w_{kl}(t+1)x_l(t), \quad (1)$$

where f_k is a differentiable (usually semi-linear) function. For all w_{ij} and for all $l \in U, t \geq 0$ we define

$$q_{ij}^l(t) = \frac{\partial net_l(t)}{\partial w_{ij}} = \sum_{\tau=1}^t \frac{\partial net_l(t)}{\partial w_{ij}(\tau)}.$$

Furthermore we define

$$e_k(t) = d_k(t) - x_k(t) \quad \text{if } k \in T(t) \text{ and } 0 \text{ otherwise,}$$

²Pineda has described another recurrent net algorithm which, as he states, “has some of the worst features of both algorithms” (Pineda, 1990). His algorithm requires $\geq O(n^4)$ memory and $\geq O(n^4)$ computations per time step, if the number of time steps exceeds n .

$$E(t) = \frac{1}{2} \sum_{k \in U} (e_k(t))^2, \quad E^{total}(t', t) = \sum_{\tau=t'+1}^t E(\tau).$$

The algorithm is a cross between the BPTT and the RTRL-algorithm. The description of the algorithm will be interleaved with its derivation and some comments concerning complexity. The basic idea is: Decompose the calculation of the gradient into blocks, each covering $O(n)$ time steps. For each block perform $n + 1$ BPTT-like passes, one pass for calculating error derivatives, and n passes for calculating derivatives of the net-inputs to the n non-input units at the end of each block. Perform $n + 1$ RTRL-like calculations for integrating the results of these BPTT-like passes into the results obtained from previous blocks.

The algorithm starts by setting the variable $t_0 \leftarrow 0$. t_0 represents the beginning of the current block. Note that for all possible $l, w_{ij} : q_{ij}^l(0) = 0, \frac{\partial E^{total}(0,0)}{\partial w_{ij}} = 0$. The main loop of the algorithm consists of 5 steps.

STEP1: Set $h \leftarrow O(n)$ (I recommend: $h \leftarrow n$).

The quantity $\frac{\partial E^{total}(0, t_0)}{\partial w_{ij}}$ for all w_{ij} is already known and $q_{ij}^l(t_0)$ is known for all appropriate l, i, j . There is an efficient way of computing the contribution of $E^{total}(0, t_0 + h)$ to the change in $w_{ij}, \Delta w_{ij}(t_0 + h)$:

$$\Delta w_{ij}(t_0 + h) = -\alpha \frac{\partial E^{total}(0, t_0 + h)}{\partial w_{ij}} = -\alpha \sum_{\tau=1}^{t_0+h} \frac{\partial E^{total}(0, t_0 + h)}{\partial w_{ij}(\tau)},$$

where α is a learning rate constant.

STEP2: Let the network run from time step t_0 to time step $t_0 + h$ according to the activation dynamics specified in equation (1). If it turns out that the current training sequence has less than $t_0 + h$ time steps (i.e., $h > s - t_0$) then $h \leftarrow s - t_0$. If $h = 0$ then EXIT.

STEP3: Perform a combination of a BPTT-like phase with an RTRL-like calculation for computing error derivatives as described next. We write

$$\begin{aligned} \frac{\partial E^{total}(0, t_0 + h)}{\partial w_{ij}} &= \frac{\partial E^{total}(0, t_0)}{\partial w_{ij}} + \frac{\partial E^{total}(t_0, t_0 + h)}{\partial w_{ij}} = \frac{\partial E^{total}(0, t_0)}{\partial w_{ij}} + \sum_{\tau=1}^{t_0+h} \frac{\partial E^{total}(t_0, t_0 + h)}{\partial w_{ij}(\tau)} \\ &= \frac{\partial E^{total}(0, t_0)}{\partial w_{ij}} + \sum_{\tau=1}^{t_0} \frac{\partial E^{total}(t_0, t_0 + h)}{\partial w_{ij}(\tau)} + \sum_{\tau=t_0+1}^{t_0+h} \frac{\partial E^{total}(t_0, t_0 + h)}{\partial w_{ij}(\tau)} \quad (2) \end{aligned}$$

The first term of (2) is already known. Consider the third term:

$$\sum_{\tau=t_0+1}^{t_0+h} \frac{\partial E^{total}(t_0, t_0 + h)}{\partial w_{ij}(\tau)} = - \sum_{\tau=t_0+1}^{t_0+h} \delta_i(\tau) x_j(\tau - 1)$$

where

$$\delta_i(\tau) = -\frac{\partial E^{total}(t_0, t_0 + h)}{\partial net_i(\tau)}.$$

For a given t_0 , $\delta_i(\tau)$ can be computed for all $i \in U, t_0 \leq \tau \leq t_0 + h$ with a single h step BPTT-pass of the order $O(hn^2)$ operations:

$$\begin{aligned} \delta_i(\tau) &= f'_i(net_i(\tau))e_i(\tau) \quad \text{if } \tau = t_0 + h \\ \delta_i(\tau) &= f'_i(net_i(\tau))(e_i(\tau) + \sum_{l \in U} w_{li}\delta_l(\tau + 1)) \quad \text{if } t_0 \leq \tau < t_0 + h \end{aligned}$$

What remains is the computation of the second term of (2) for all w_{ij} , which requires $O(n^3)$ operations:

$$\begin{aligned} \sum_{\tau=1}^{t_0} \frac{\partial E^{total}(t_0, t_0 + h)}{\partial w_{ij}(\tau)} &= \sum_{\tau=1}^{t_0} \sum_{k \in U} \frac{\partial E^{total}(t_0, t_0 + h)}{\partial net_k(t_0)} \frac{\partial net_k(t_0)}{\partial w_{ij}(\tau)} \\ &= \sum_{k \in U} -\delta_k(t_0) \sum_{\tau=1}^{t_0} \frac{\partial net_k(t_0)}{\partial w_{ij}(\tau)} = -\sum_{k \in U} \delta_k(t_0) q_{ij}^k(t_0). \end{aligned}$$

STEP4: To compute $q_{ij}^l(t_0 + h)$ for all possible l, i, j , perform n combinations of a BPTT-like phase with an RTRL-like calculation (one such combination for each l) for computing as follows:

$$\begin{aligned} q_{ij}^l(t_0 + h) &= \frac{\partial net_l(t_0 + h)}{\partial w_{ij}} = \sum_{\tau=1}^{t_0+h} \frac{\partial net_l(t_0 + h)}{\partial w_{ij}(\tau)} = \sum_{\tau=1}^{t_0} \frac{\partial net_l(t_0 + h)}{\partial w_{ij}(\tau)} + \sum_{\tau=t_0+1}^{t_0+h} \frac{\partial net_l(t_0 + h)}{\partial w_{ij}(\tau)} \\ &= \sum_{\tau=1}^{t_0} \sum_{k \in U} \frac{\partial net_l(t_0 + h)}{\partial net_k(t_0)} \frac{\partial net_k(t_0)}{\partial w_{ij}(\tau)} + \sum_{\tau=t_0+1}^{t_0+h} \frac{\partial net_l(t_0 + h)}{\partial net_i(\tau)} \frac{\partial net_i(\tau)}{\partial w_{ij}(\tau)} \\ &= \sum_{k \in U} \gamma_{lk}(t_0) h_{ij}^k(t_0) + \sum_{\tau=t_0+1}^{t_0+h} \gamma_{li}(\tau) x_j(\tau - 1) \end{aligned} \quad (3)$$

where

$$\gamma_{lk}(\tau) = \frac{\partial net_l(t_0 + h)}{\partial net_k(\tau)}.$$

For a given t_0 , a given $l \in U$ and for all $i \in U, t_0 \leq \tau \leq t_0 + h$ the quantity $\gamma_{li}(\tau)$ can be computed with a single h step BPTT-like operation of the order $O(hn^2)$:

$$\begin{aligned} \text{if } \tau = t_0 + h: \quad &\text{if } l = i \text{ then } \gamma_{li}(\tau) = 1 \text{ else } \gamma_{li}(\tau) = 0 \\ \text{if } t_0 \leq \tau < t_0 + h: \quad &\gamma_{li}(\tau) = f'_i(net_i(\tau)) \sum_{a \in U} w_{ai} \gamma_{la}(\tau + 1) \end{aligned}$$

For a given l , the computation of (3) for all w_{ij} requires $O(n^3 + hn^2)$ operations. Therefore STEP3 and STEP4 together require $(n + 1)O(hn^2 + n^3)$ operations *spread over h time steps*. Since $h = O(n)$, $O(n^4)$ computations are spread over $O(n)$ time steps. *This implies an average of $O(n^3)$ computations per time step.*

The final step of the algorithm's main loop is

STEP5: Set $t_0 \leftarrow t_0 + h$ and go to STEP1.

The off-line version of the algorithm waits until the end of an episode (which needs not be known in advance) before performing weight changes. An on-line version performs weight changes each time STEP4 is completed.

As formulated above, the algorithm needs $O(n^4)$ computations at its peak, every n -th time step. Nothing prevents us, however, from distributing these $O(n^4)$ computations more evenly over n time steps. One way of achieving this is to perform one of the n BPTT-like phases of STEP 4 at each time step of the next 'block' of n time steps.

CONCLUDING REMARKS

Like the RTRL-algorithm the method needs a fixed amount of storage of the order $O(n^3)$. Like the RTRL-algorithm (but unlike the methods described in (Williams and Peng, 1990) and (Zipser, 1989)) the algorithm computes the exact gradient. Since it is $O(n)$ times faster than RTRL, it should be preferred.

Following the argumentation in (Williams and Peng, 1990), continuous time versions of BPTT and RTRL (Pearlmutter, 1989) (Gherry, 1989) can serve as a basis for a correspondingly efficient continuous time version of the algorithm presented here (by means of Euler discretization).

Many typical environments produce input sequences that have both local and more global temporal structure. For instance, input sequences are often hierarchically organized (e.g. speech). In such cases, sequence-composing algorithms (Schmidhuber, 1991) (Schmidhuber, 1992) can provide superior alternatives to pure gradient-based algorithms.

1 ACKNOWLEDGEMENTS

Thanks to Mike Mozer, Bernd Schürmann, and Daniel Prelinger for providing useful comments on an earlier draft of this paper.

References

- Gherrity, M. (1989). A learning algorithm for analog fully recurrent neural networks. In *IEEE/INNS International Joint Conference on Neural Networks, San Diego*, volume 1, pages 643–644.
- Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, 1(2):263–269.
- Pineda, F. J. (1990). Time dependent adaptive neural networks. In Touretzky, D. S., editor, *Advances in Neural Information Processing Systems 2*, pages 710–718. San Mateo, CA: Morgan Kaufmann.
- Robinson, A. J. and Fallside, F. (1987). The utility driven dynamic error propagation network. Technical Report CUED/F-INFENG/TR.1, Cambridge University Engineering Department.
- Schmidhuber, J. H. (1991). Adaptive decomposition of time. In Kohonen, T., Mäkisara, K., Simula, O., and Kangas, J., editors, *Artificial Neural Networks*, pages 909–914. Elsevier Science Publishers B.V., North-Holland.
- Schmidhuber, J. H. (1992). Learning complex, extended sequences using the principle of history compression. *Neural Computation*, 4(2):234–242.
- Williams, R. J. (1989). Complexity of exact gradient computation algorithms for recurrent neural networks. Technical Report Technical Report NU-CCS-89-27, Boston: Northeastern University, College of Computer Science.
- Williams, R. J. and Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation*, 4:491–501.
- Williams, R. J. and Zipser, D. (1989). Experimental analysis of the real-time recurrent learning algorithm. *Connection Science*, 1(1):87–111.
- Williams, R. J. and Zipser, D. (1992). Gradient-based learning algorithms for recurrent networks and their computational complexity. In *Back-propagation: Theory, Architectures and Applications*. Hillsdale, NJ: Erlbaum.
- Zipser, D. (1989). A subgrouping strategy that reduces learning complexity and speeds up learning in recurrent networks. *Neural Computation*, 1(4):552–558.